Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 9.1: Fusion, co-learning and new trends

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Lecture Objectives

- Quick recap: multimodal fusion
- Model-agnostic fusion
  - Multimodal fusion architecture search
- Fusion and kernel function
  - Transformers through the lens for kernel
  - Multiple Kernel Learning
- Co-learning
  - Paired and weakly-paired data
- Research trends in Multimodal ML NEW papers
  - Few-shot and weakly supervised learning
  - Multi-lingual multimodal grounding

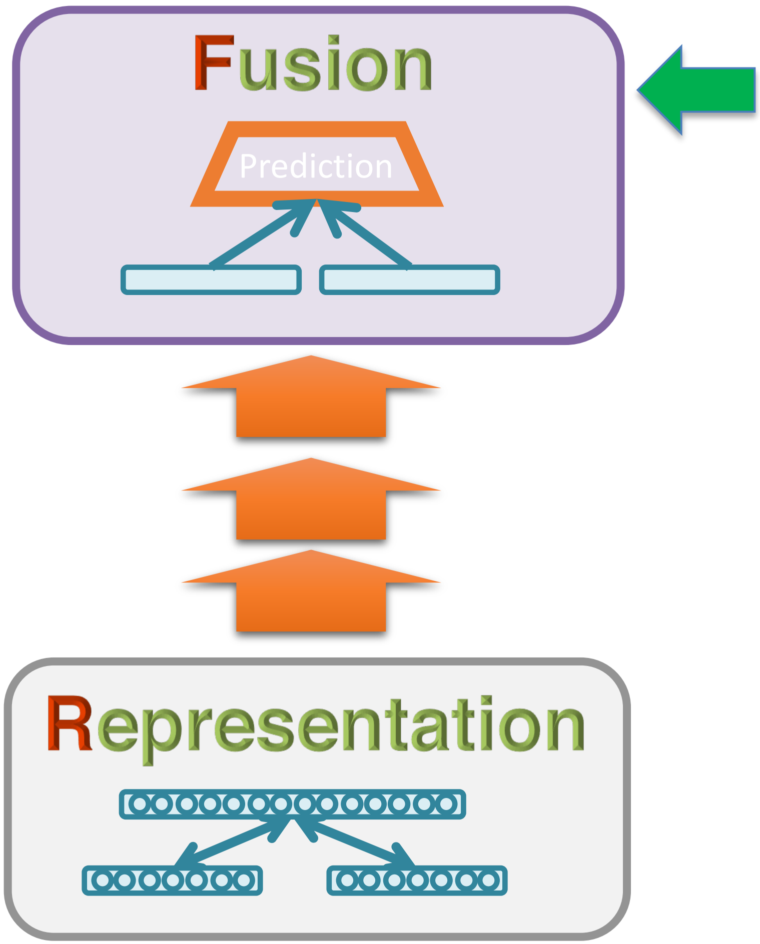
Quick Recap: Multimodal Fusion

Multimodal fusion

- Process of joining information from two or more modalities to perform a prediction
- Examples
  - Audio-visual speech recognition
  - Audio-visual emotion recognition
  - Multimodal biometrics
  - Speaker identification and diarization
  - Visual/Media Question answering

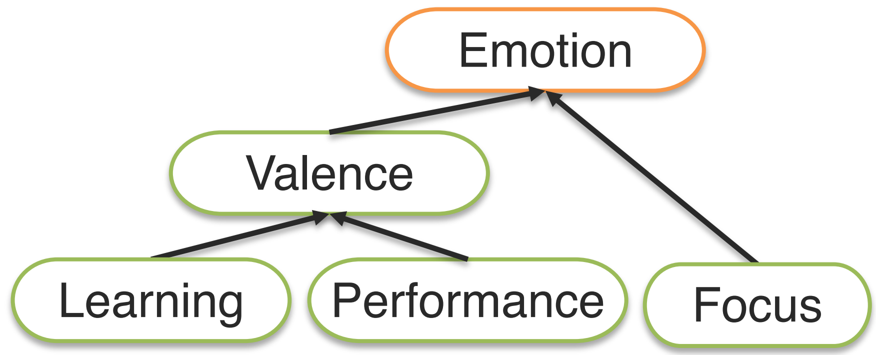
(a) answer-phone (a) get-out-car (a) fight-person

Fusion – Probabilistic Graphical Models

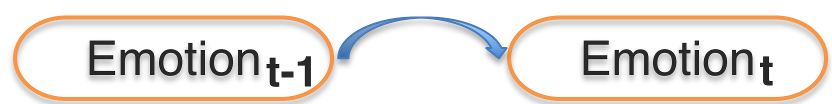


Domain knowledge

a) Latent sub-structure



b) Structured output prediction



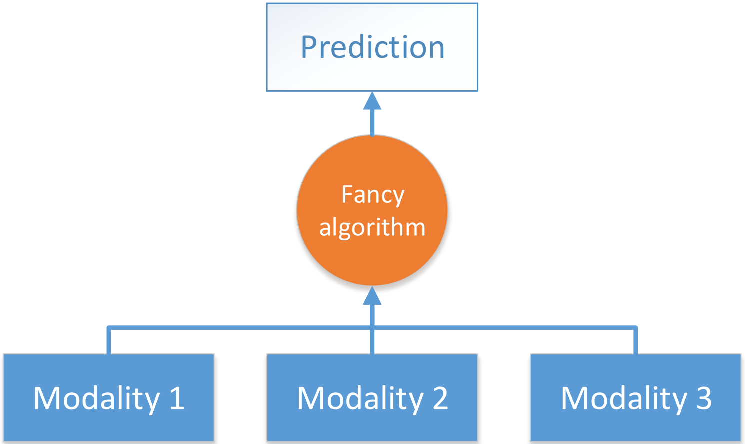
Multimodal Fusion

"Model-agnostic" fusion:

- Early and late fusion
- Fusion architecture search

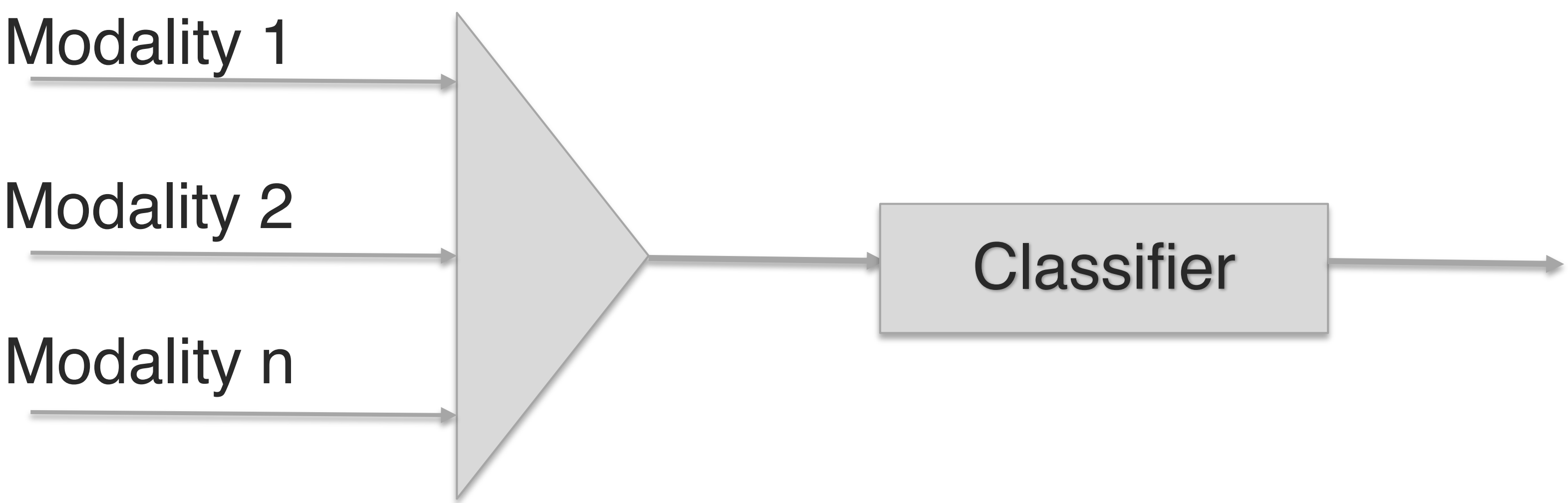
Intermediate fusion (aka model-based):

- Neural Networks
- Graphical models
- Kernel Methods



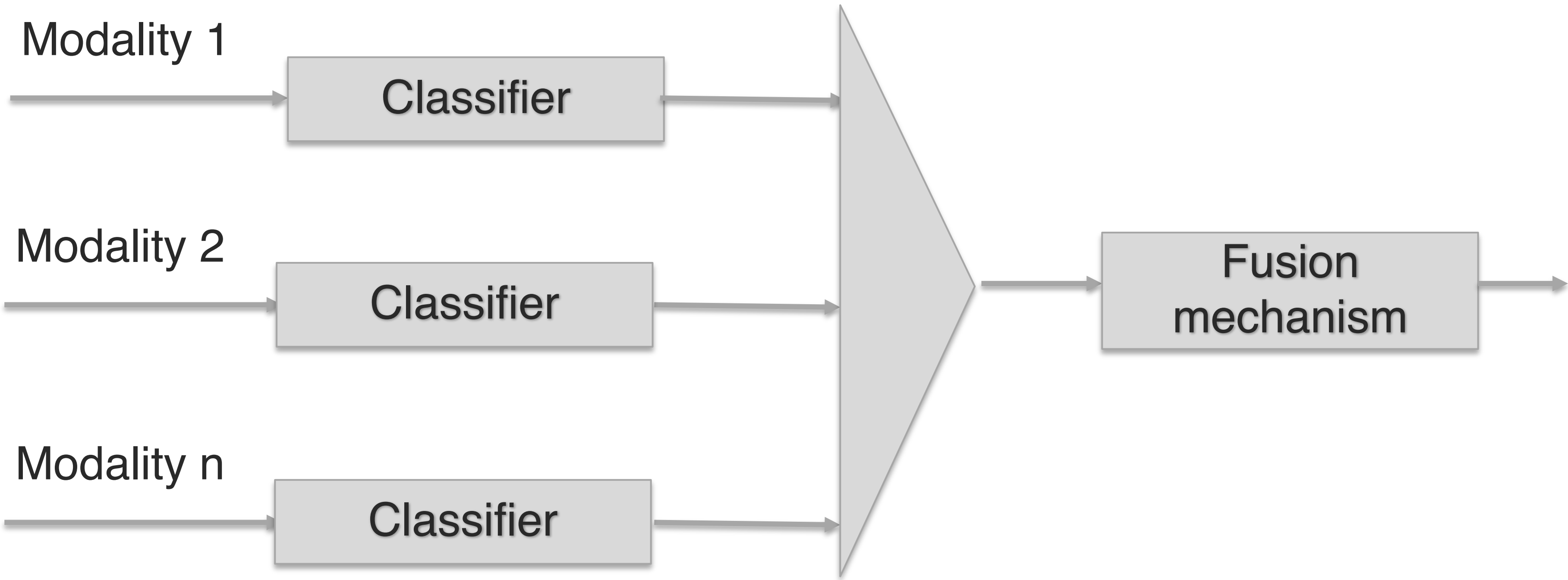
Model-free Fusion

Model-agnostic approaches – early fusion



- Easy to implement – just concatenate the features
- Exploit dependencies between features
- Can end up very high dimensional
- More difficult to use if features have different granularities

Model-agnostic approaches – late fusion



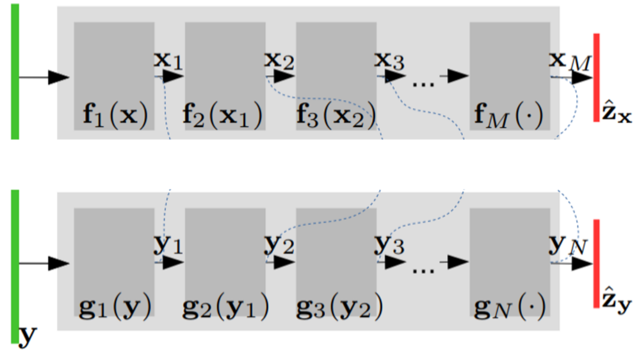
- Train a unimodal predictor and a multimodal fusion one
- Requires multiple training stages
- Do not model low level interactions between modalities
- Fus … pproach

What should be the Fusion Mechanism for multi-layer neural classifiers?

Late Fusion on Multi-Layer Unimodal Classifiers

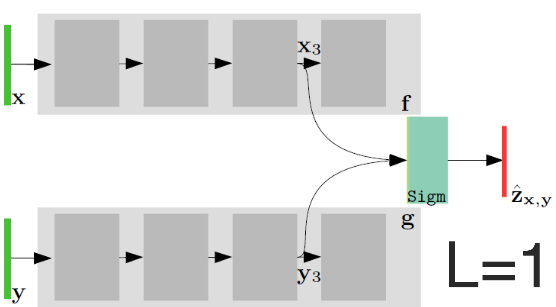
Unimodal classifier 1

Unimodal classifier 2

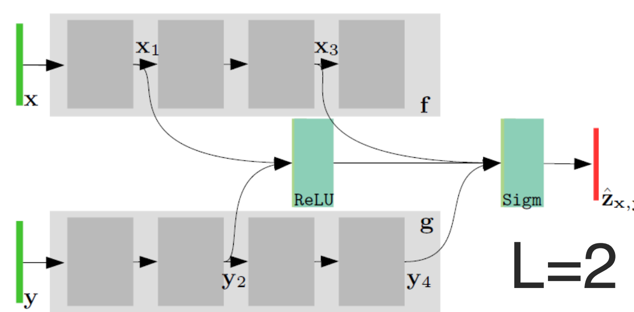


What layer(s) should we fuse?

One of the last layers?



Or more than one layer?



Trying all combinations may be computationally expensive…

Multimodal Fusion Architecture Search (MFAS)

NEW-ish paper

Proposed solution: Explore the search space with *Sequential Model-Based Optimization*

- Start with simpler models first (all L=1 models) and iteratively increase the complexity (L=2, L=3,…)
- Use a *surrogate* function to predict performance of unseen architectures
  - e.g., the performance of all the L=1 models should give us an idea of how well the L=2 models will perform

“Perez-Rua, Vielzeuf, Pateux, Baccouche, Frederic Jurie,MFAS: Multimodal Fusion Architecture Search, CVPR 2019

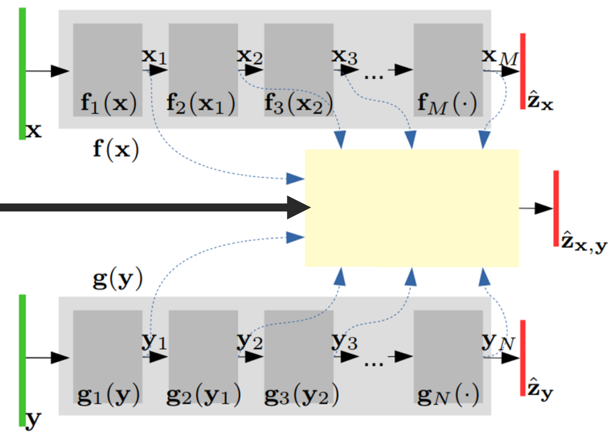
Multimodal Fusion Architecture Search (MFAS)

Basic building block: a “fusion layer” unit

Unimodal classifier 1

Fusion layer unit

Unimodal classifier 2



With three hyper-parameters:

a) Layer index for modality 1
b) Layer index for modality 2
c) Activation/fusion function

“Perez-Rua, Vielzeuf, Pateux, Baccouche, Frederic Jurie,MFAS: Multimodal Fusion Architecture Search, CVPR 2019

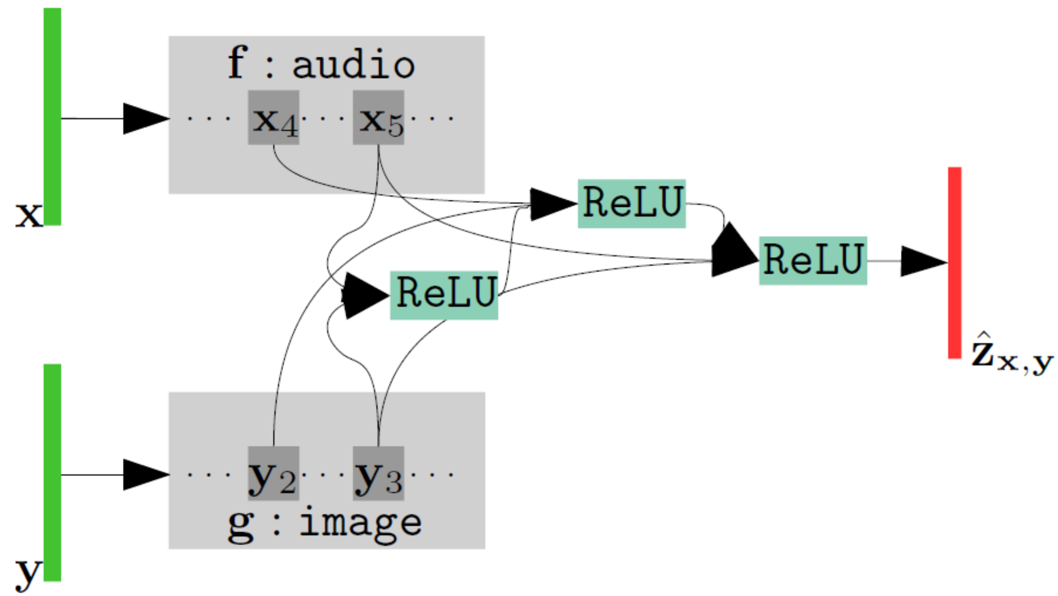
Multimodal Fusion Architecture Search (MFAS)

Dataset: Audio-Visual MNIST

Example of discovered fusion architecture with MFAS:

LeNet trained on spoken digits

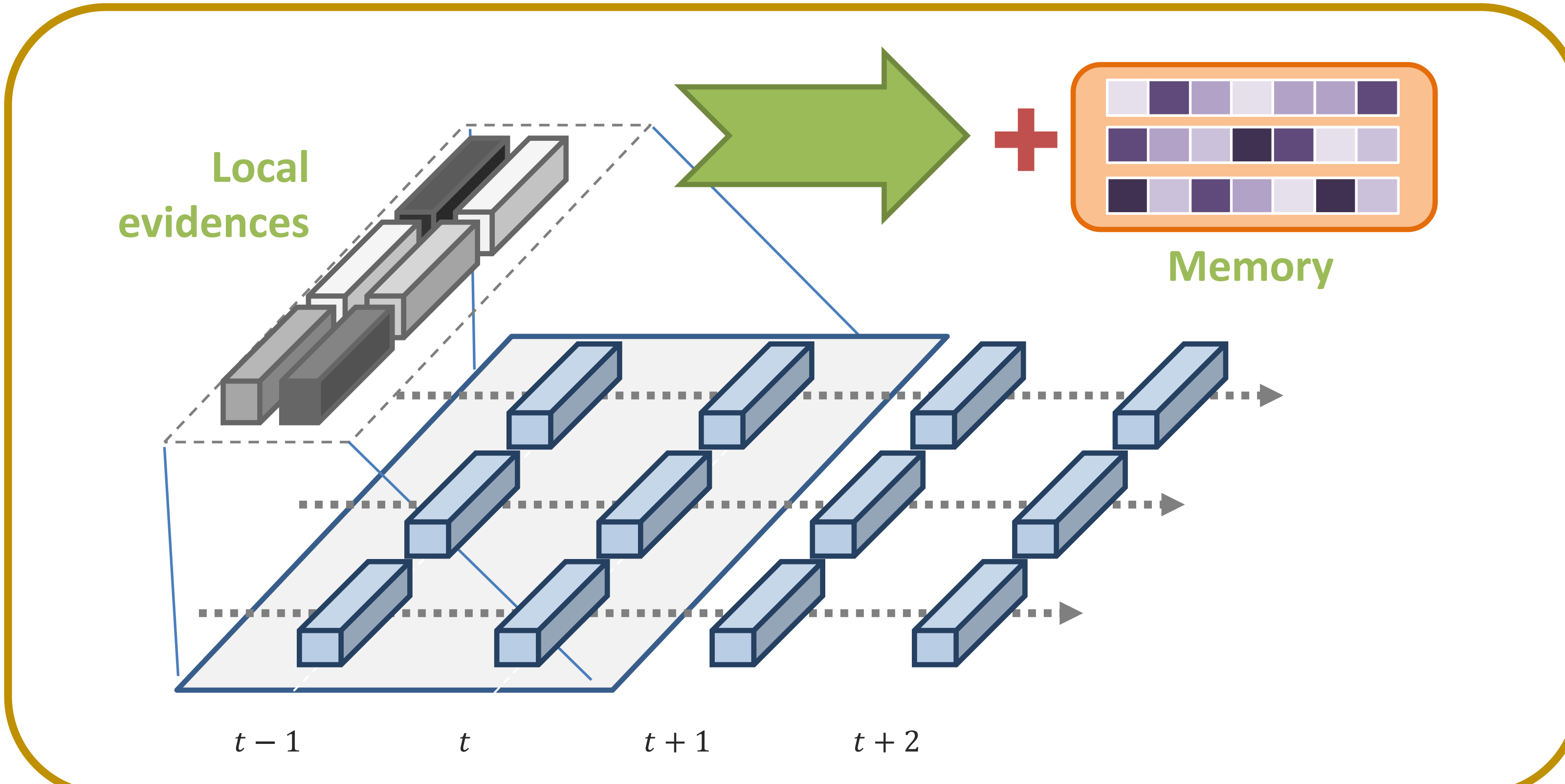
LeNet trained on visual digits



“Perez-Rua, Vielz

What should be the Fusion Mechanism for variable length unimodal classifier?

Memory-Based Fusion

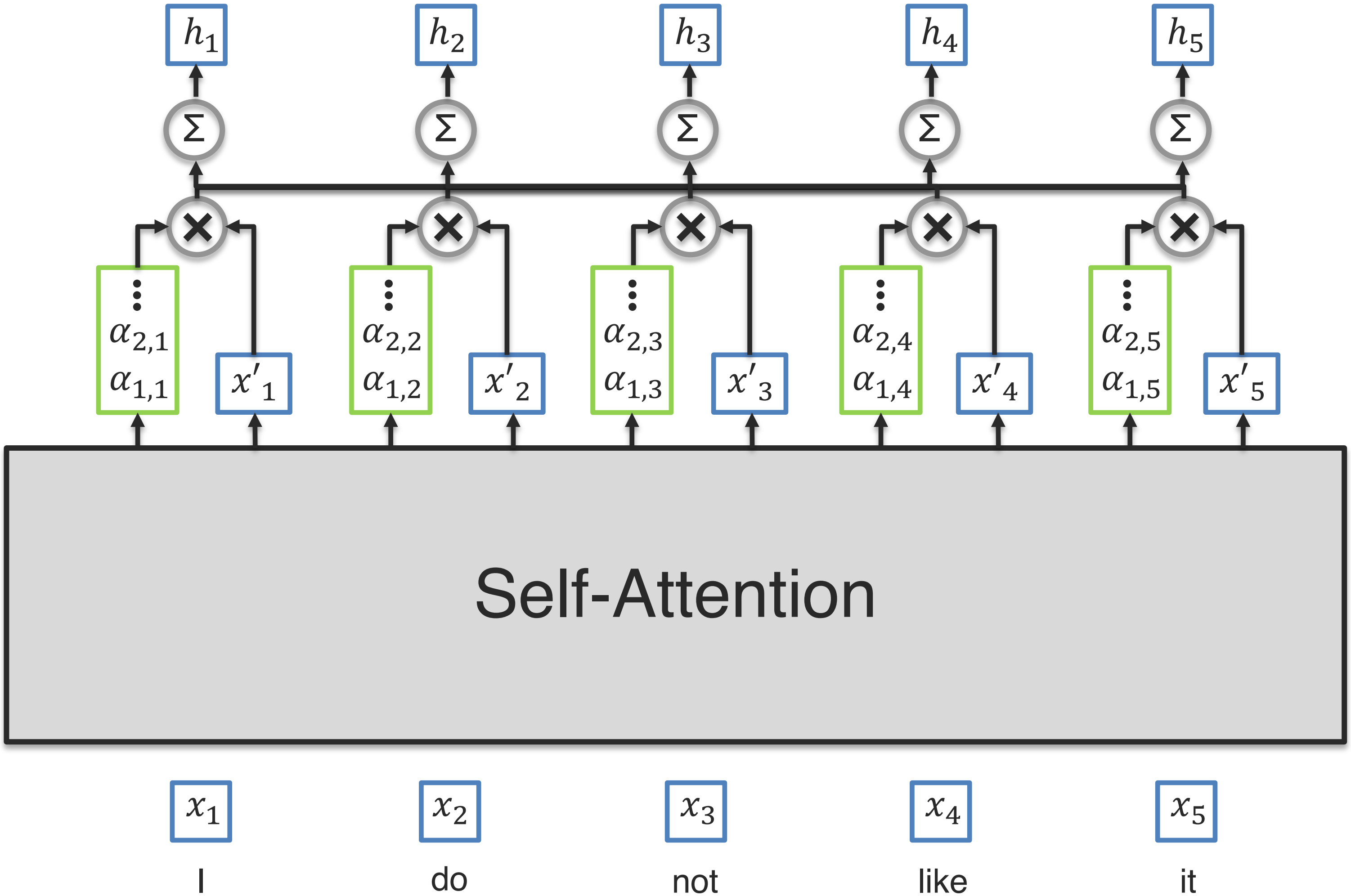


➢ This model can also be trained end-to-end.

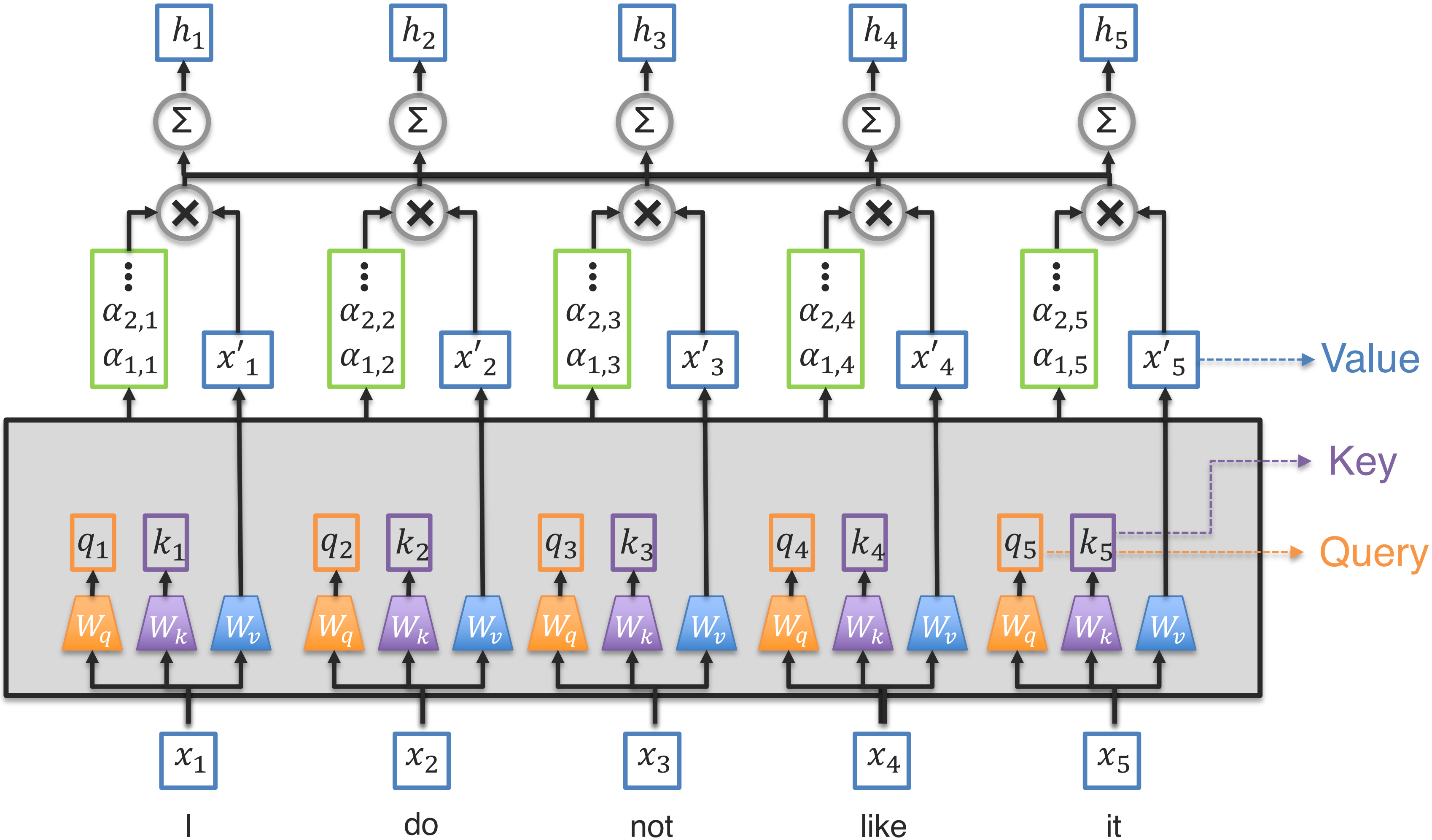
[Zadeh et al., Memory Fusion Network for Multi-view Sequential Learning, AAAI 2018]

Local Fusion and Kernel Functions

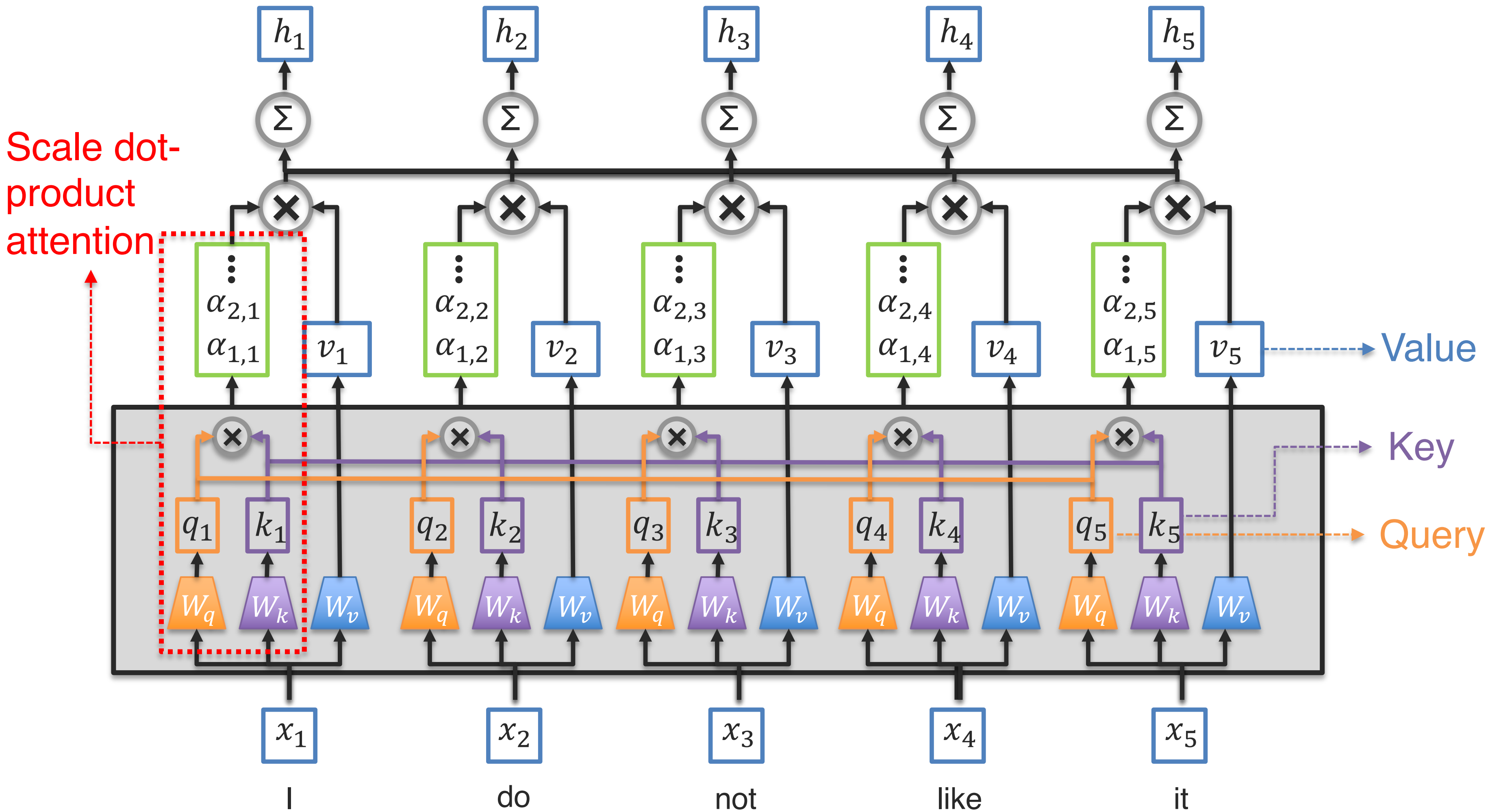
Recap: Self-Attention



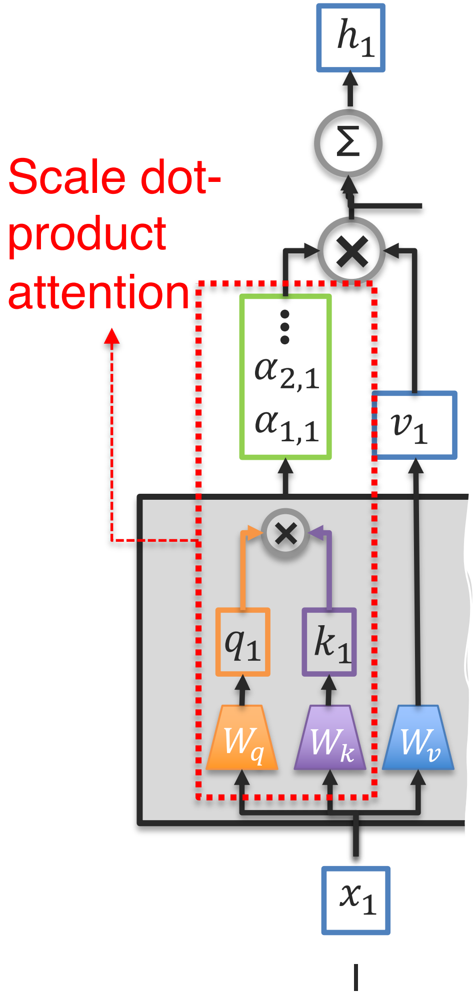
Recap: Transformer Self-Attention



Transformer Self-Attention



Transformer's Attention Function



Scale dot-product attention:

$$\boldsymbol{\alpha} = softmax\left(\frac{\boldsymbol{x}_q \boldsymbol{W}_q (\boldsymbol{x}_k \boldsymbol{W}_k)^T}{\sqrt{d_k}}\right)$$

This attention function is a similarity function. This is related to kernel function…

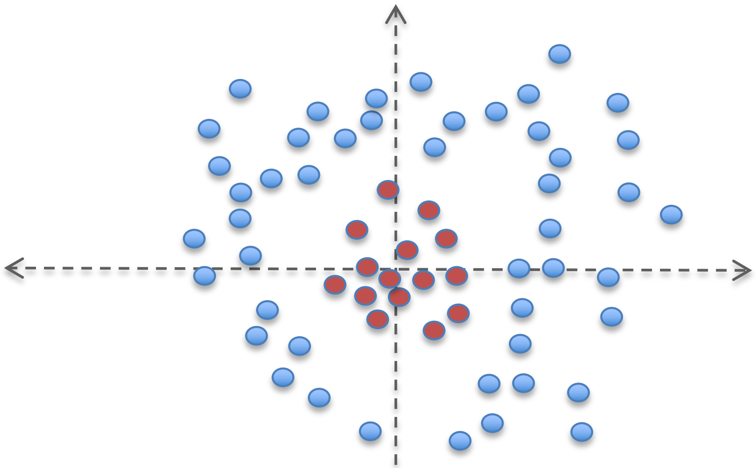
What is a Kernel function?

A kernel function: Acts as a similarity metric between data points

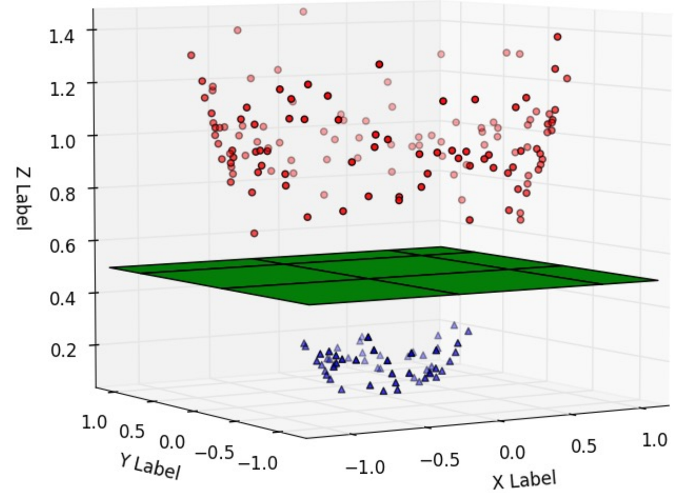
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle, \text{ where } \phi: D \rightarrow Z$$

- Kernel function performs an inner product in feature map space ϕ
- Inner product (a generalization of the dot product) is often denoted as $\langle .,. \rangle$ in SVM papers
- $\boldsymbol{x} \in \mathbb{R}^D$(but not necessarily), but $\phi(\boldsymbol{x})$ can be in any space – same, higher, lower or even in an infinite dimensional space

Non-linearly separable data



Not linearly separable



Same data, but now linearly separable

- Want to map our data to a linearly separable space
- Instead of $\boldsymbol{x}$, want $\phi(\boldsymbol{x})$, in a separable space ($\phi(\boldsymbol{x})$ is a feature map)

What if $\phi(\boldsymbol{x})$ is much higher dimensional? We do not want to learn more parameters and mapping could become very expensive

Radial Basis Function Kernel (RBF)

Arguably the most popular kernel function (for Support Vector Machine)

$$K(x_i, x_j) = \exp - \frac{1}{2\sigma^2} \left\| x_i - x_j \right\|^2$$

$\phi(\boldsymbol{x}) = ?$

- It is infinite dimensional and fairly involved, no easy way to actually perform the mapping to this space, but we know what an inner product looks like in it

$\sigma = ?$

- a hyperparameter
- With a really low sigma the model becomes close to a KNN approach (potentially very expensive)

Some other kernels

Other kernels exist

- Histogram Intersection Kernel
  - good for histogram features
- String kernels
  - specifically for text and sentence features
- Proximity distribution kernel
- (Spatial) pyramid matching kernel

Kernel CCA

If we remember CCA it used only inner products in definitions when dealing with data, that means we can again use kernels

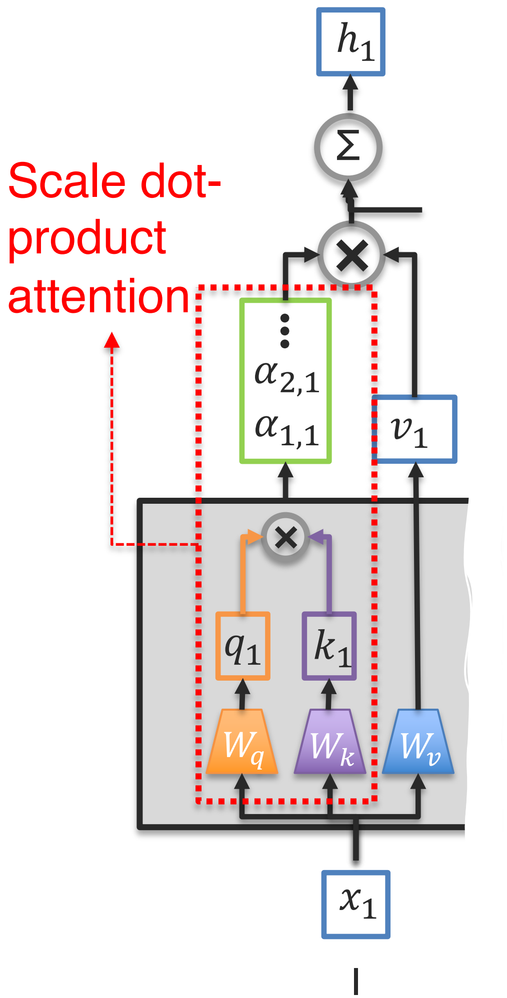
$$(w_1^*, w_2^*) = \underset{w_1, w_2}{\operatorname{argmax}} \frac{w_1' \Sigma_{12} w_2}{\sqrt{w_1' \Sigma_{11} w_1 w_2' \Sigma_{22} w_2}} = \underset{w_1' \Sigma_{11} w_1 = w_2' \Sigma_{22} w_2 = 1}{\operatorname{argmax}} w_1' \Sigma_{12} w_2$$

We can now map into a high-dimensional non-linear space instead

$$(\alpha_1^*, \alpha_2^*) = \underset{\alpha_1, \alpha_2}{\operatorname{argmax}} \frac{\alpha_1' K_1 K_2 \alpha_2}{\sqrt{(\alpha_1' K_1^2 \alpha_2)\ (\alpha_1' K_2^2 \alpha_2)}} = \underset{\alpha_1' K_1^2 \alpha_1 = \alpha_2' K_2^2 \alpha_2 = 1}{\operatorname{argmax}} \alpha_1' K_1 K_2 \alpha_2,$$

[Lai et al. 2000]

Transformer's Attention Function

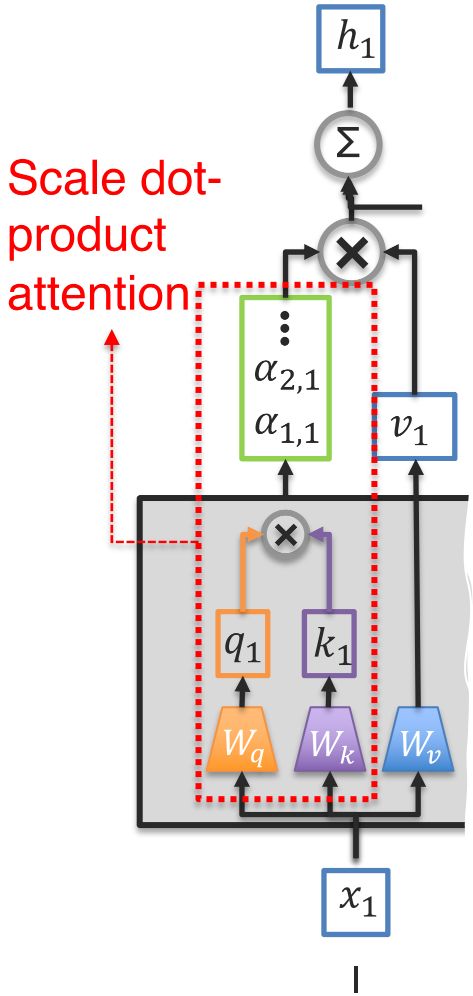


Scale dot-product attention:

$$\boldsymbol{\alpha} = softmax\left(\frac{\boldsymbol{x_q W_q}(\boldsymbol{x_k W_k})^T}{\sqrt{d_k}}\right)$$

How can you interpret it as a kernel similarity function?

Transformer's Attention Function



Scale dot-product attention:

$$\boldsymbol{\alpha} = softmax\left(\frac{\boldsymbol{x}_q \boldsymbol{W}_q (\boldsymbol{x}_k \boldsymbol{W}_k)^T}{\sqrt{d_k}}\right)$$

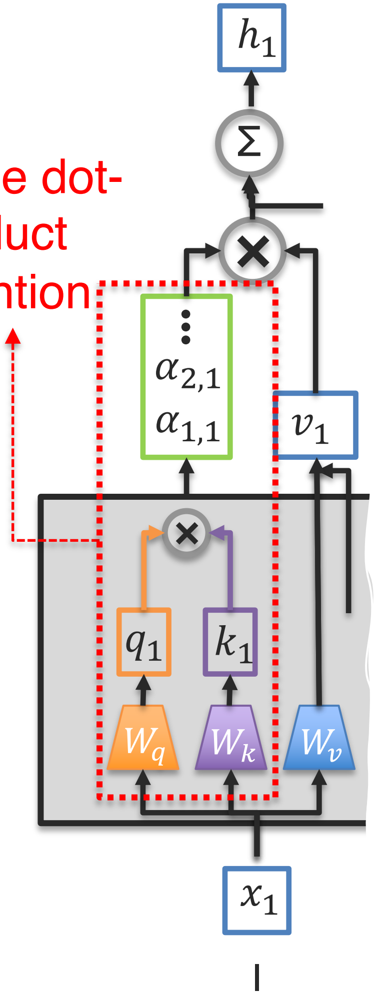
Kernel-formulated attention:

$$\boldsymbol{\alpha} = \frac{k(\boldsymbol{x}_q, \boldsymbol{x}_k)}{\sum_{\{\boldsymbol{x}'_k\}} k(\boldsymbol{x}_q, \boldsymbol{x}'_k)}$$

What is the impact of the kernel function?

Tsai et al., Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel, EMNLP 2019

Transformer's Attention Function



What is the impact of the kernel function?

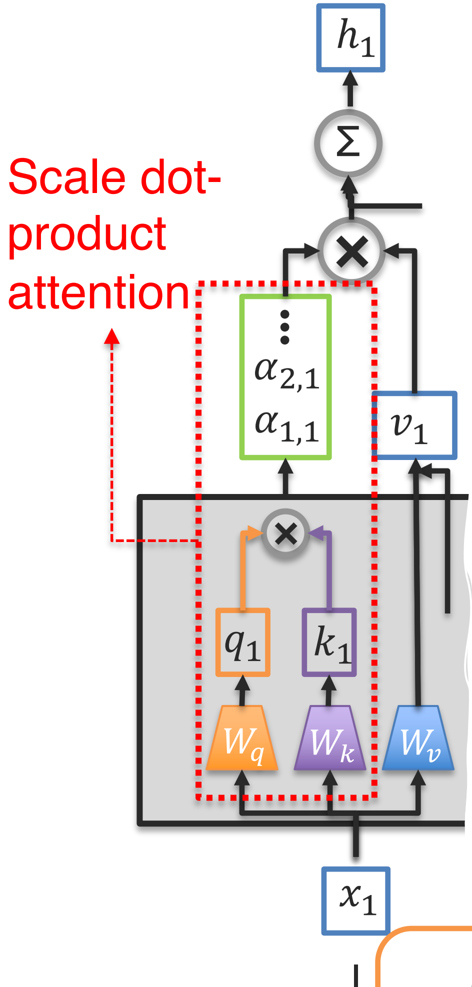
Type	Kernel Form	NMT (BLEU↑) Asym. ($W_q \neq W_k$)	NMT (BLEU↑) Sym. ($W_q = W_k$)
Linear	$\langle f_a W_q, f_b W_k \rangle$	not converge	not converge
Polynomial	$\left(\langle f_a W_q, f_b W_k \rangle\right)^2$	32.72	32.43
Exponential	$\exp\left(\frac{\langle f_a W_q, f_b W_k \rangle}{\sqrt{d_k}}\right)$	33.98	33.78
RBF	$\exp\left(-\frac{\lVert f_a W_q - f_b W_k \rVert^2}{\sqrt{d_k}}\right)$	**34.26**	34.14

Conventional Transformer → Exponential

What is the best way to integrate the position embedding?

Tsai et al., Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel, EMNLP 2019

Transformer's Attention Function



What is the best way to integrate the position embedding?

PE Incorporation	Kernel Form	NMT (BLEU↑)
Vaswami et al → Direct-Sum	$k_{\exp}\left(f_q + t_q, f_k + t_k\right)$	33.98
Look-up Table	$L_{t_q - t_k, f_q} \cdot k_{\exp}\left(f_q, f_k\right)$	34.12
Transformer XL → Product Kernel	$k_{\exp}\left(f_q, f_k\right) \cdot k_{f_q}\left(t_q, t_k\right)$	33.62
Proposed → Product Kernel	$k_F\left(f_q, f_k\right) \cdot k_T\left(t_q, t_k\right)$	**34.71**

$$\text{with } k_F(f_q, f_k) = \exp\left(\frac{\langle f_q W_F, f_k W_F \rangle}{\sqrt{d_k}}\right)$$

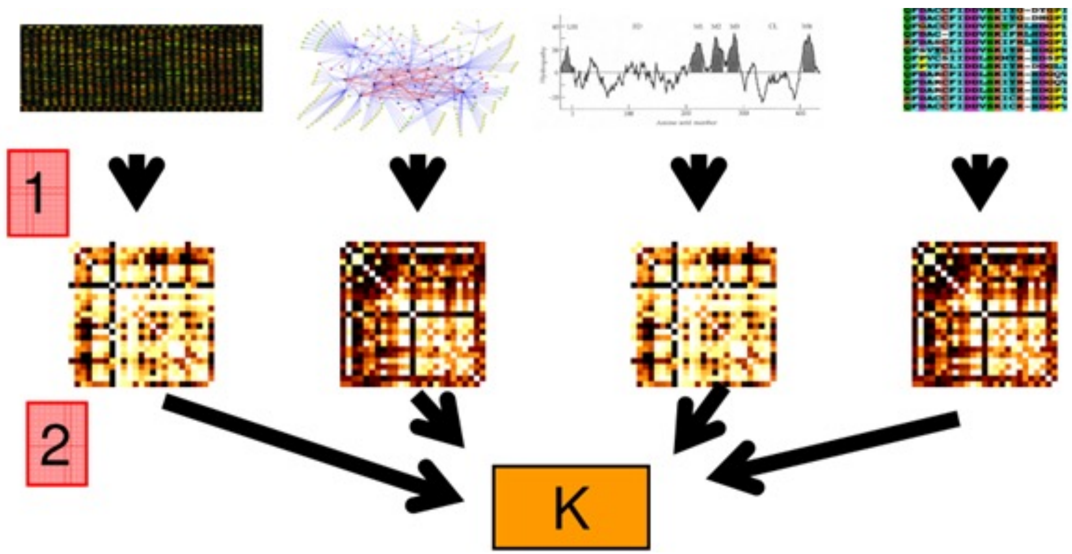
$$\text{and } k_T(t_q, t_k) = \exp\left(\frac{\langle t_q W_T, t_k W_T \rangle}{\sqrt{d_k}}\right),$$

Same weight matrices!

ansformer's

Can Kernels be used as a Fusion Mechanism (for late fusion)?

Multiple Kernel Learning

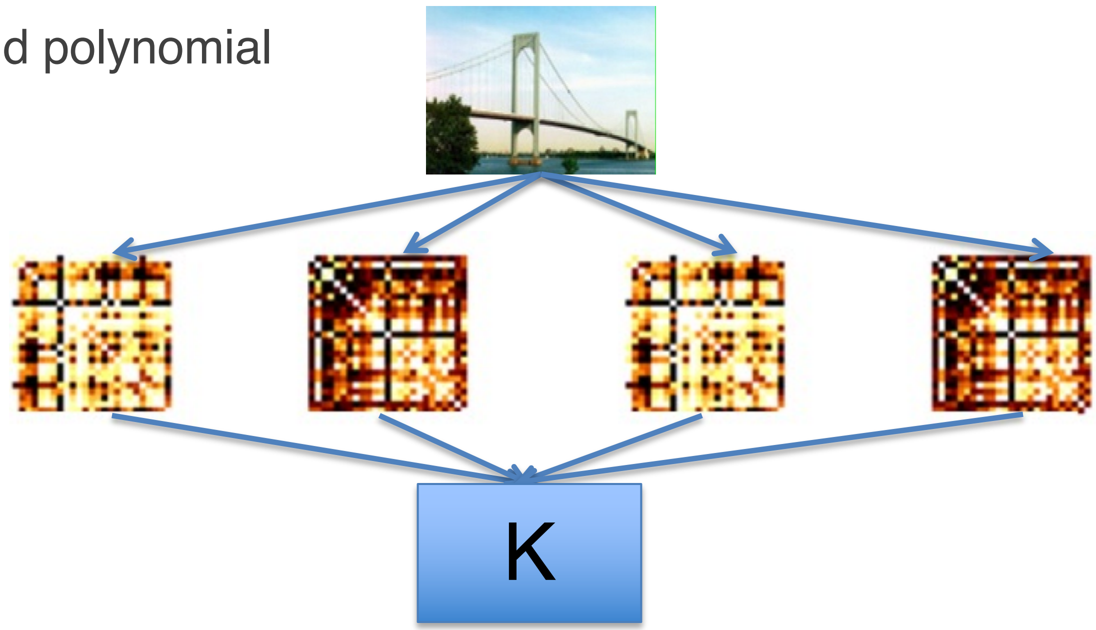


[Lanckriet 2004]

- Instead of providing a single kernel and validating which one works optimize in a family of kernels (or different families for different modalities)
- Works well for unimodal and multimodal data, very little adaptation is needed

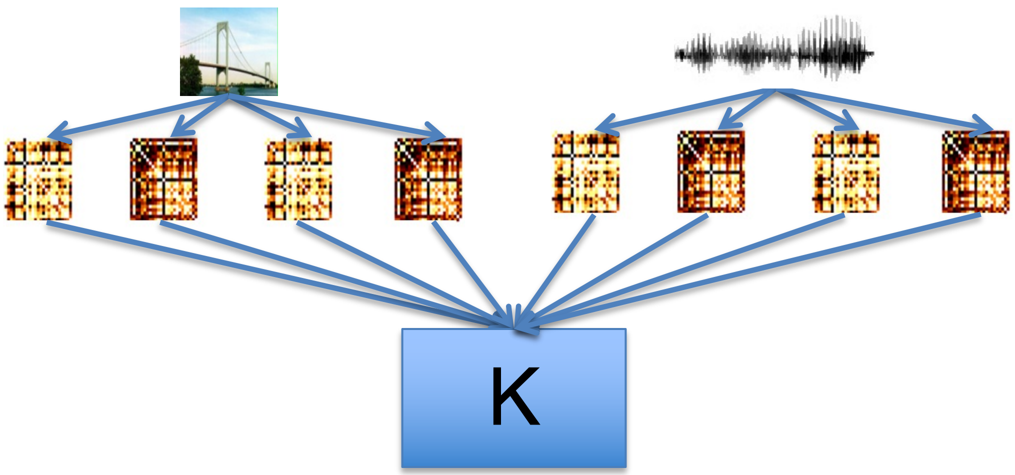
MKL in Unimodal Case

- Pick a family of kernels and learn which kernels are important for the classification case
- For example a set of RBF and polynomial kernels



MKL in Multimodal/Multiview Case

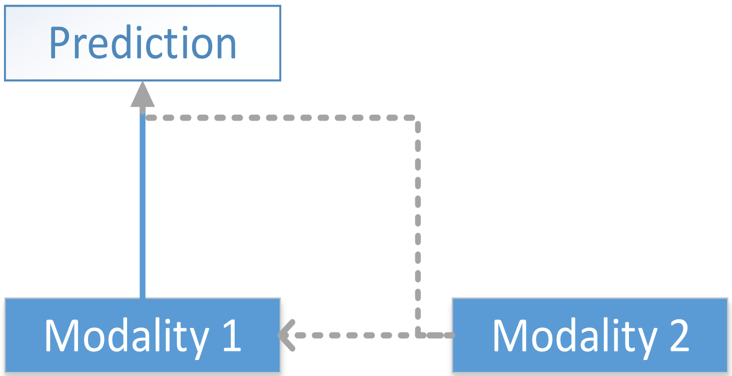
- Pick a family of kernels for each modality and learn which kernels are important for the classification case
- Does not need to be different modalities, often we use different views of the same modality (HOG, SIFT, etc.)

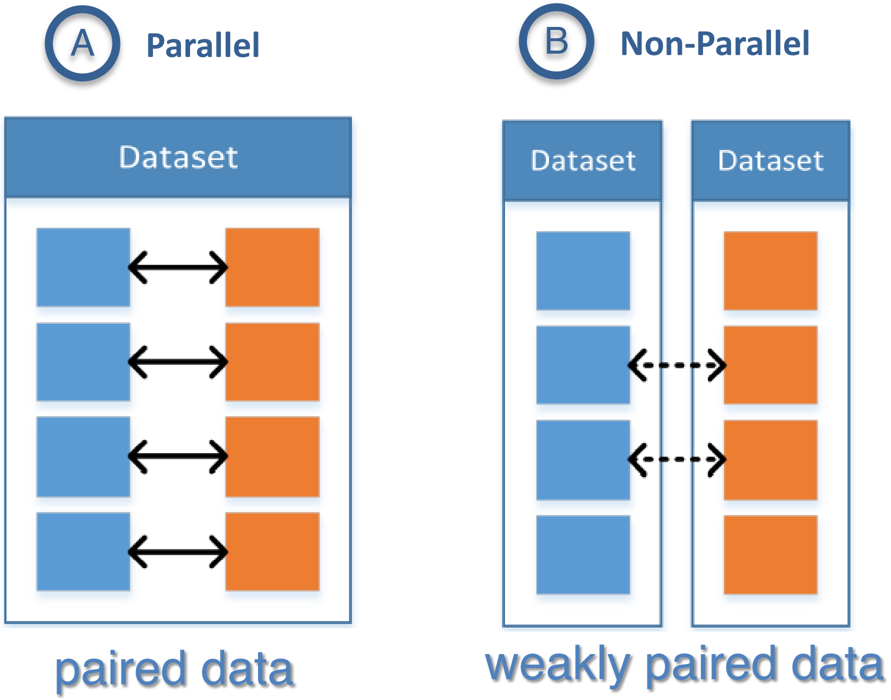


Co-Learning

Co-Learning - The 5th Multimodal Challenge

Definition: Transfer knowledge between modalities, including their representations and predictive models.



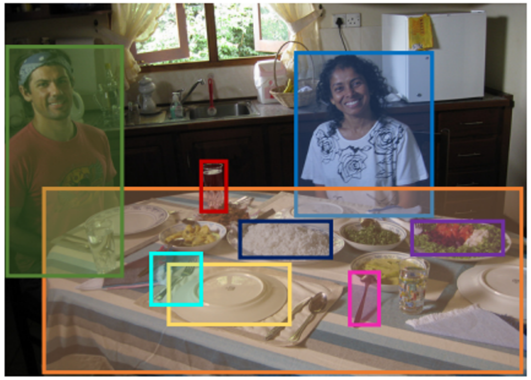


Co-learning Example with Paired Data

Learn vector representations for text using visual co-occurrences

Four types of co-occurrences:

(a) Object - Attribute
(b) Attribute - Attribute
(c) Context
(d) Object-Hypernym



Region	Object Words	Attribute Words
	man, person, adult, mammal	muscular, smiling
	woman, person, adult, mammal	lean, smiling
	table, tablecloth, furniture	striped, oval
	rice, carbohydrates, food	white, grainy, cooked
	salad, roughage, food	leafy, chopped, healthy, red, green
	glass, glassware, utensil	clear, transparent, reflective, tall
	plate, crockery, utensil	ceramic, white, round, circular
	fork, cutlery, utensil	metallic, shiny, reflective
	spoon, cutlery, utensil	serving, metallic, shiny, reflective

ViCo: Word Embeddings from Visual Co-occurrences

ViCo: Word Embeddings from Visual Co-occurrences

Relatedness through Co-occurrences

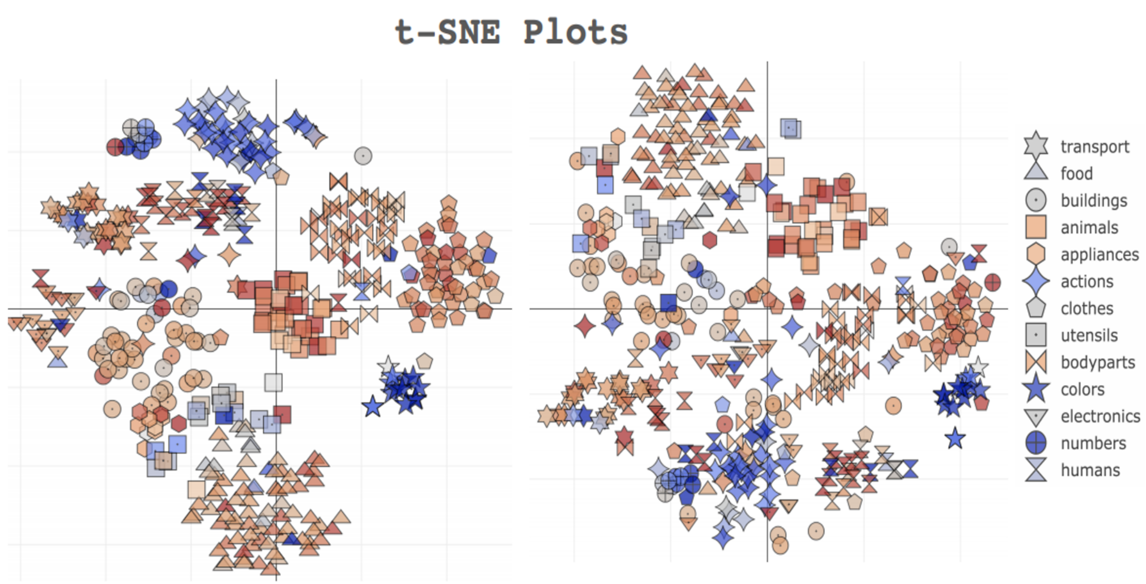
Word Pair	ViCo	Obj-Attr	Attr-Attr	Obj-Hyp	Context	GloVe
crouch / squat	0.61	0.74	0.72	0.18	0.25	0.05
sweet / dessert	0.66	0.78	0.76	0.56	0.79	0.43
man / male	0.71	0.98	0.8	0.38	1	0.34
purple / violet	0.75	0.93	1	0.24	0.03	0.52
hosiery / sock	0.52	0.27	0.18	0.87	0.07	0.23
aeroplane / aircraft	0.73	0.43	0.07	0.87	0.75	0.43
bench / pew	0.63	0.67	0.09	0.79	-0.14	0.1
keyboard / mouse	0.19	0.63	0.19	0.09	0.95	0.52
laptop / desk	0.39	0.23	0.24	0.1	0.94	0.28
window / door	0.59	0.46	0.35	0.53	0.93	0.67
hair / blonde	0.16	0.56	0.32	-0.15	0.17	0.51
thigh / ankle	0.09	0.19	0.03	0.01	0.39	0.74
garlic / onion	0.36	-0.03	0.3	0.37	0.56	0.77
driver / car	0.27	0.16	0.26	0.12	0.53	0.71
girl / boy	0.41	0.38	0.22	0.44	0.74	0.83

Since ViCo is learned from multiple types of co-occurrences, it is hypothesized to provide a richer sense of relatedness

➢ Learned using a multi-task Log-Bilinear Model

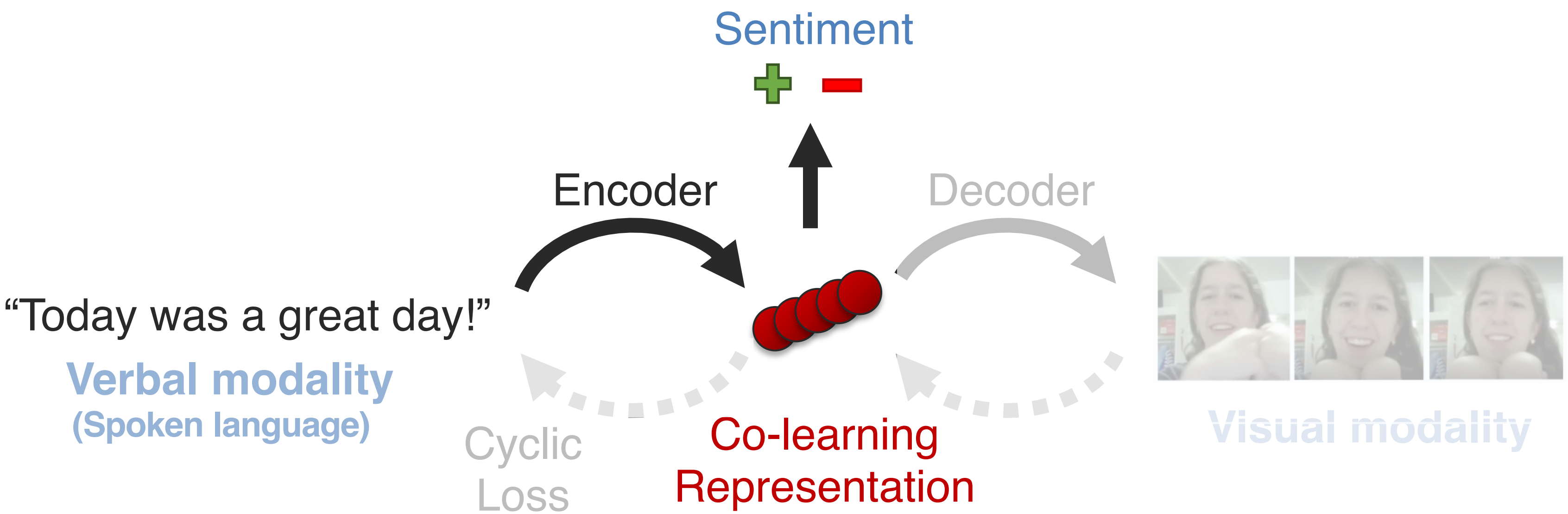
ViCo: Word Embeddings from Visual Co-occurrences

ViCO leads to more homogenous clusters compared to GloVe



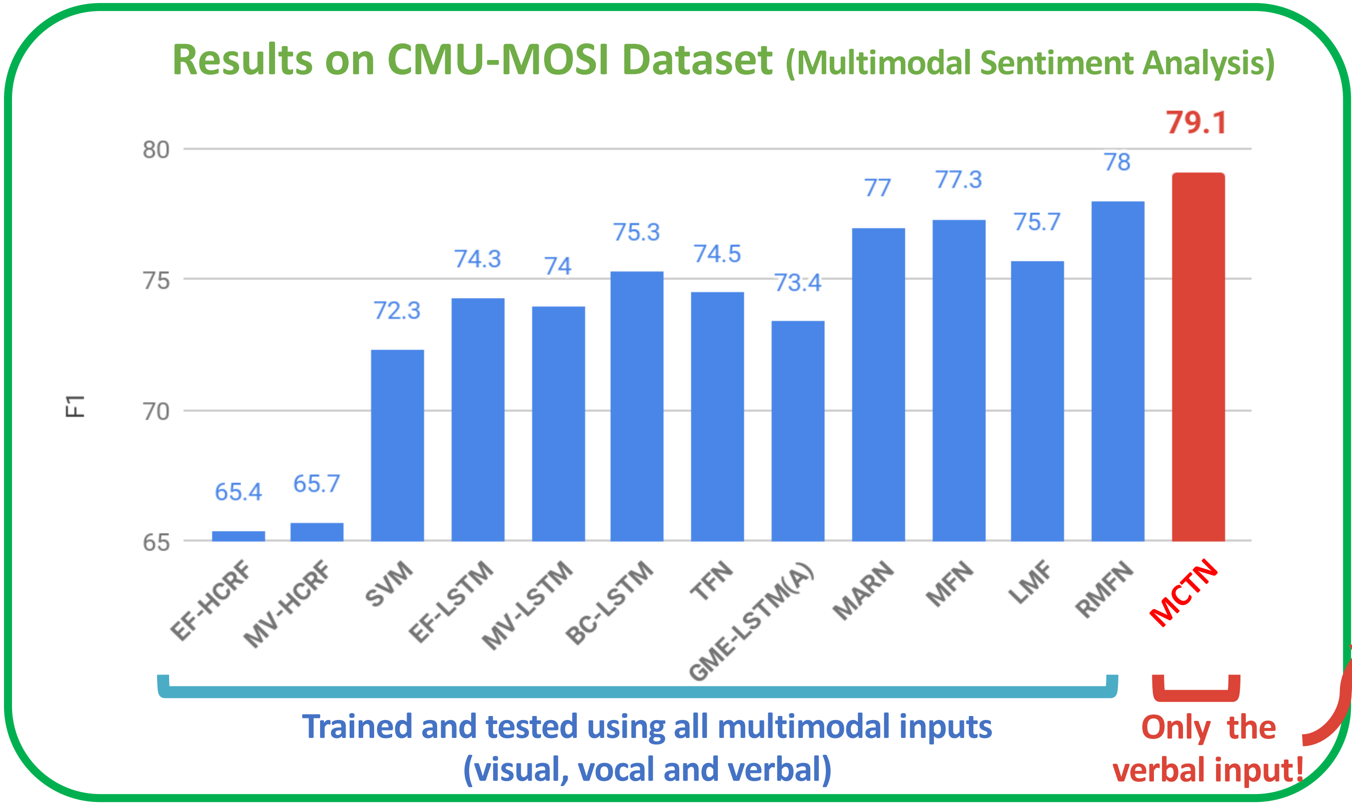
(a) GloVe+ViCo(linear) (b) GloVe

Another Example of Co-Learning with Paired Data: Multimodal Cyclic Translation



Paul Pu Liang*, Hai Pham*, et al., "Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities", AAAI 2019

Another Example of Co-Learning with Paired Data: Multimodal Cyclic Translation



Trained using multimodal co-learning but tested with only verbal input!

Paul Pu Liang*, Hai Pham*, et al., "Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities", AAAI 2019

Co-Learning Example with Weakly Paired Data

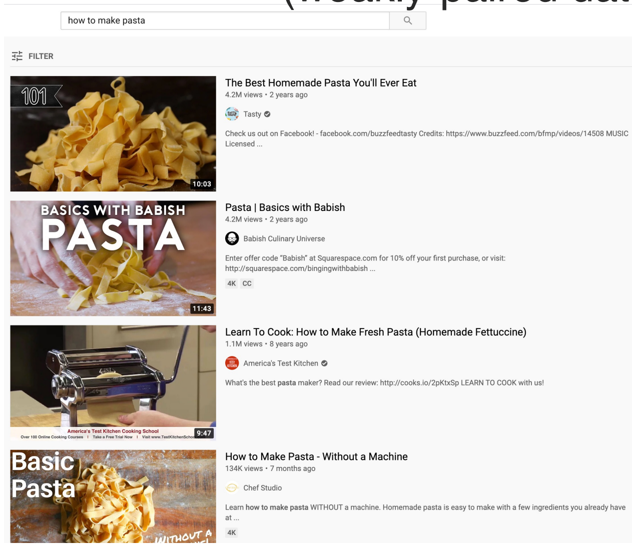
NEW paper

Goal: Learn better visual representations…

… by taking advantage of large-scale video+language resources

Instructional videos

(weakly-paired data)



it's turning into a much thicker mixture

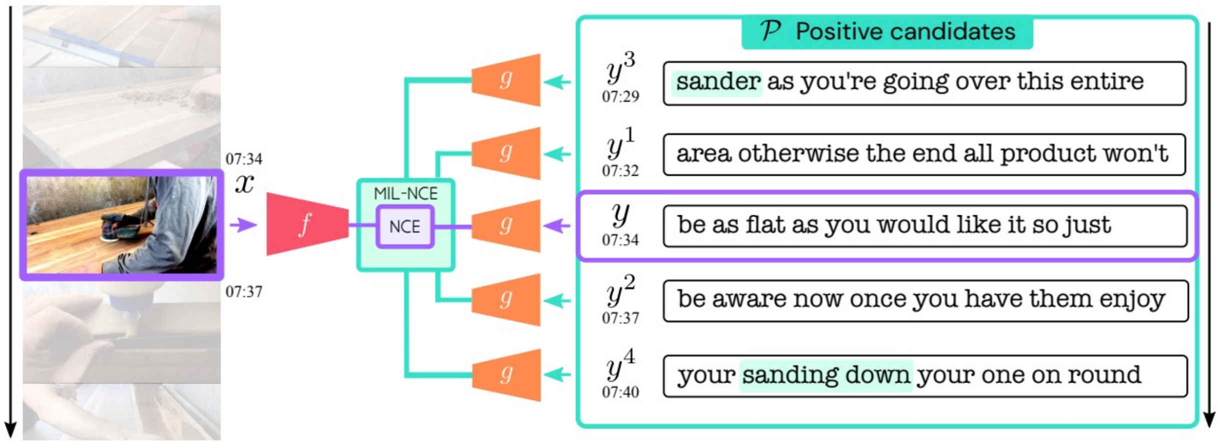
The biggest mistake is not kneading it enough

…

End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Weakly Paired Data

Data point: "a short 3.2 seconds video clip (32 frames at 10 FPS) together with a small number of words (not exceeding 16)"



How to handle this misalignment? Multi-instance learning!

How to do it self-supervised? Contrastive learning!

End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Multiple Instance Learning Noise Contrastive Estimation

Objective

Given video x and text y from a positive set P_i and a negative set N_i, maximize the positive / total score ratio

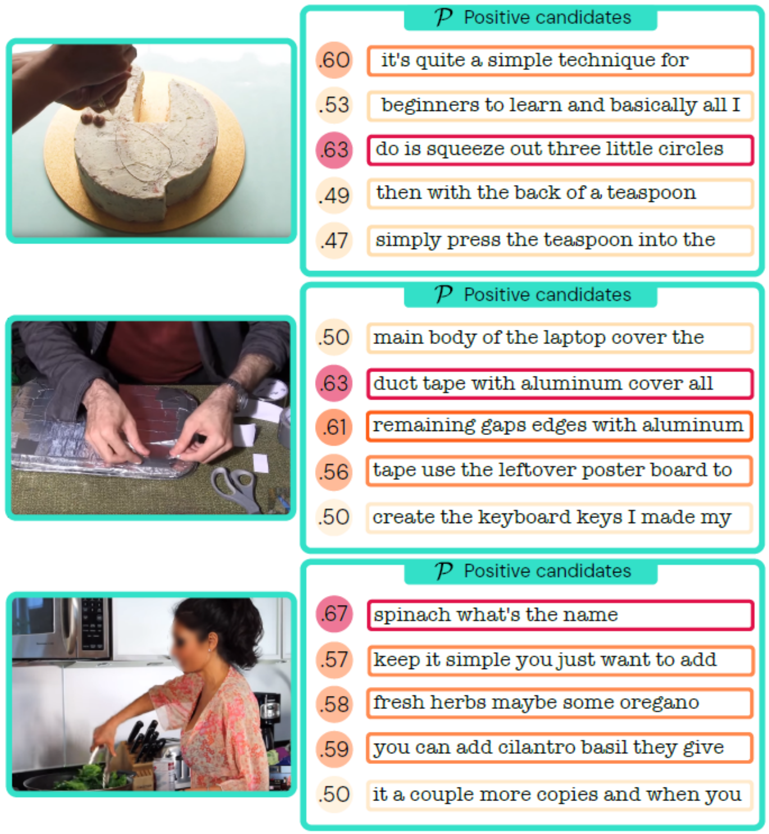
$$\max_{f,g} \sum_{i=1}^{n} \log\left(\frac{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum\limits_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum\limits_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}}\right)$$

Note: Doing so requires maximizing $f(x)^\top g(y)$ for only positive examples

1. Using sets of positive and negative examples to ~wash out the misaligned text
2. Ideally, we would maximize all positives over all possible negatives (intractable)

End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Experiments – HowTo100M Dataset



End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

New Directions

Learning by Abstraction: The Neural State Machine

NEW paper

How to solve this question using visual reasoning?

What is the red fruit inside the bowl to the right of the coffee maker?

1. Given an **image**, generate a probabilistic **scene graph** that captures the semantic concepts.

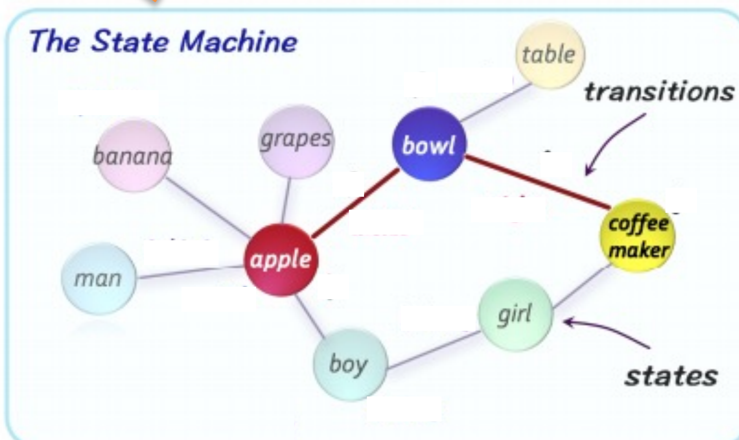
2. Treat the graph as a **state machine** and simulate iterative computation over it to *answer questions* or *draw inferences*.

3. Natural language questions are translated into *soft instructions* and used to perform sequential reasoning over the scene graph/state machine.

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine.“ NeurIPS 2019

Learning by Abstraction: The Neural State Machine

Detect objects and create proximity graph

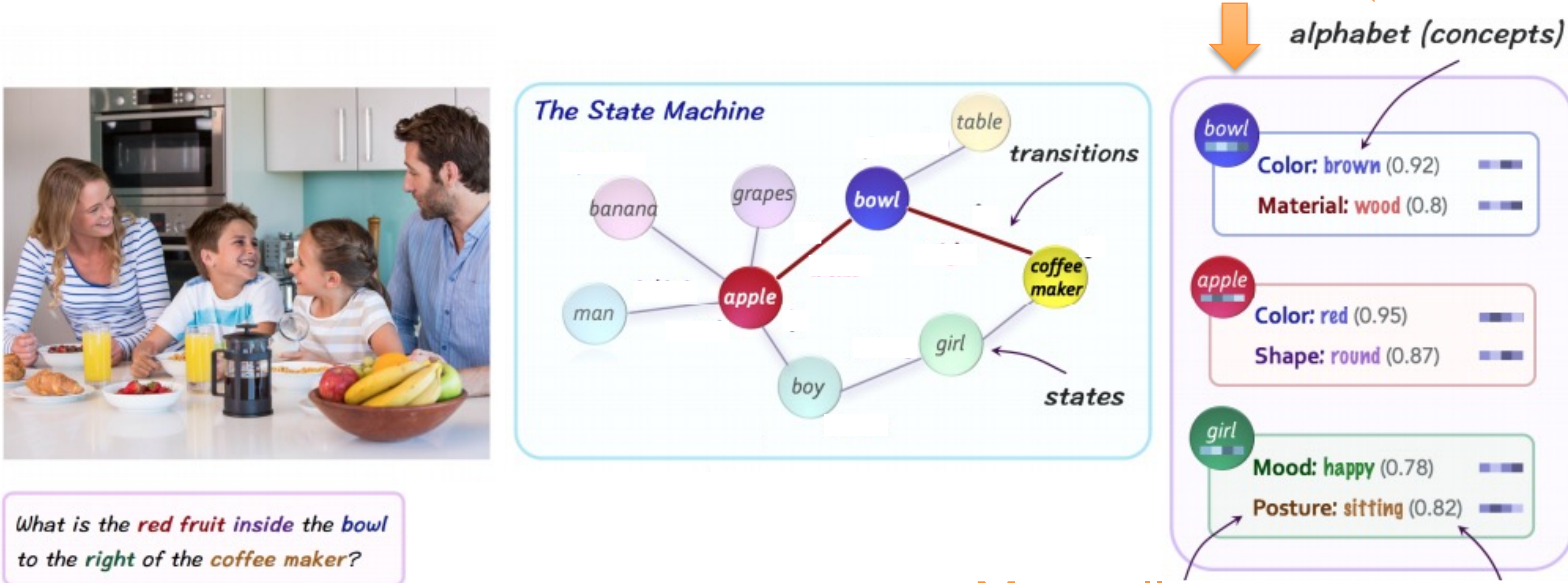


What is the red fruit inside the bowl to the right of the coffee maker?

Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

Learning by Abstraction: The Neural State Machine

Pre-trained an alphabet of c
(Visual Genome)



Manually grouped by "properties"

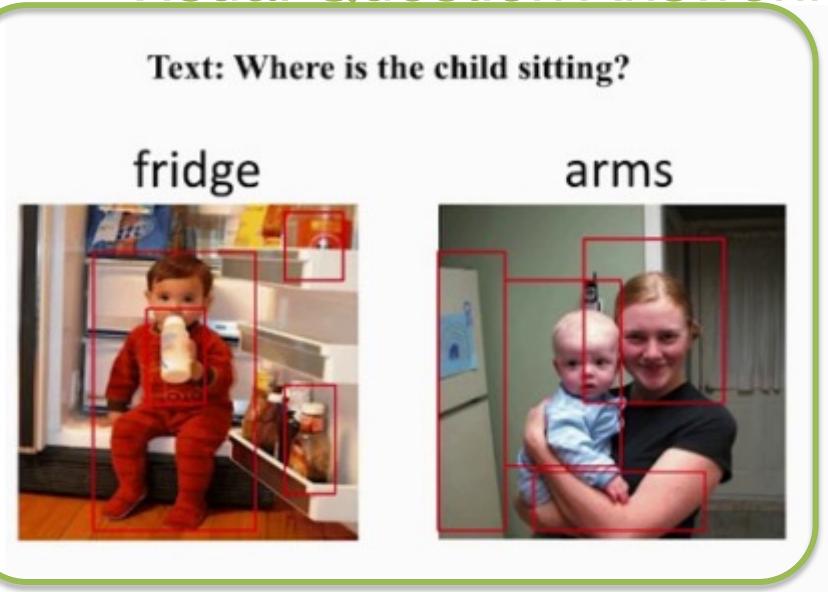
Probabilities computed at runtime for each object instance

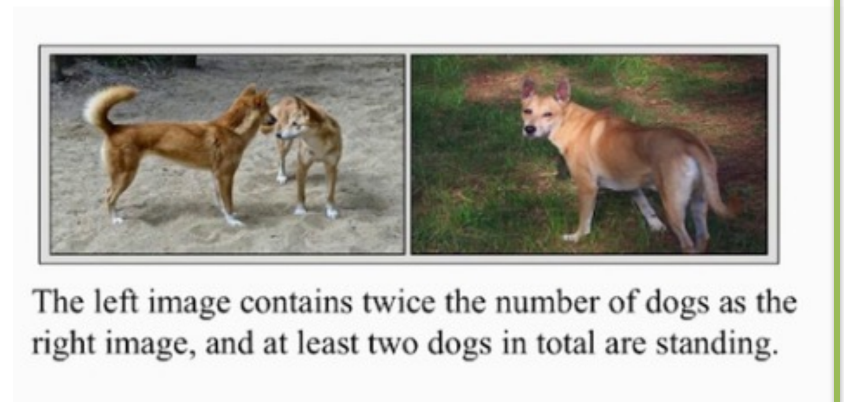
Hudson, Drew, and Christopher D. Manning. "Learning by abstraction: The neural state machine." NeurIPS 2019

Cross-Modality Relevance for Reasoning on Language and Vision

NEW paper

Visual Question AnsweringNatural Language for Visual Reas





Solving these problems requires:

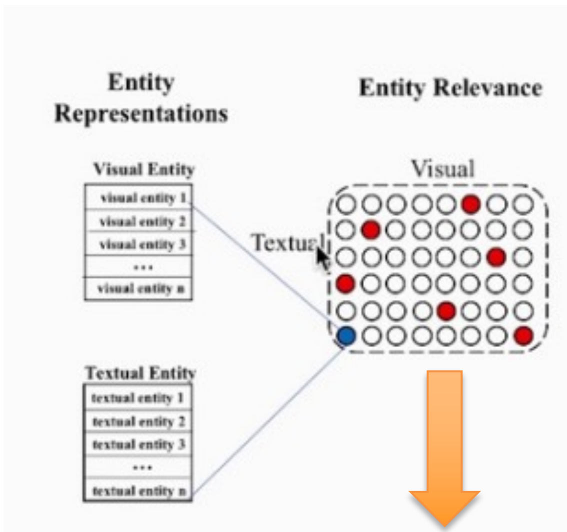
(1) Knowing relevance (aka, alignment) between visual and

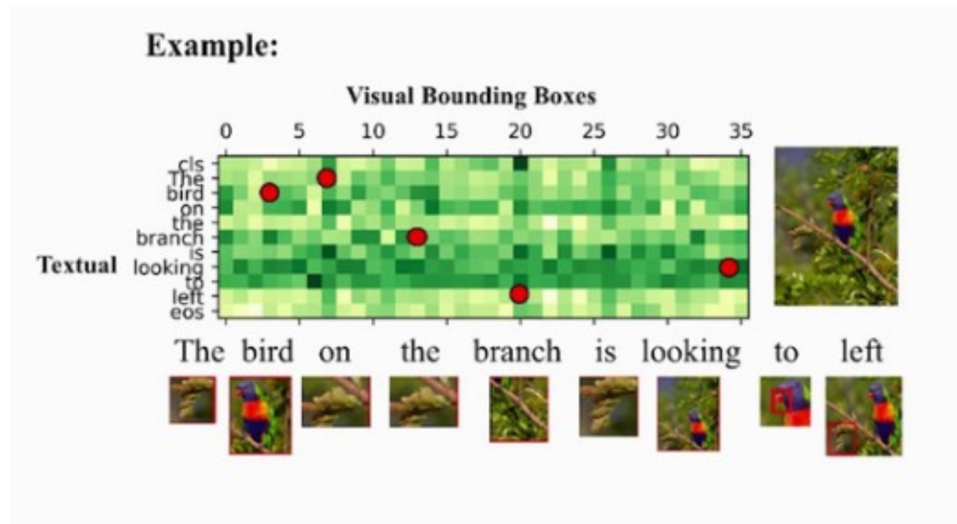
(2) Knowing relevance between visual pairs and language p

Cross-Modality Relevance for Reasoning on Language and Vision, ACL 2020

Cross-Modality Relevance for Reasoning on Language and Vision

Computing **Cross Modality Relevance** affinity matrix



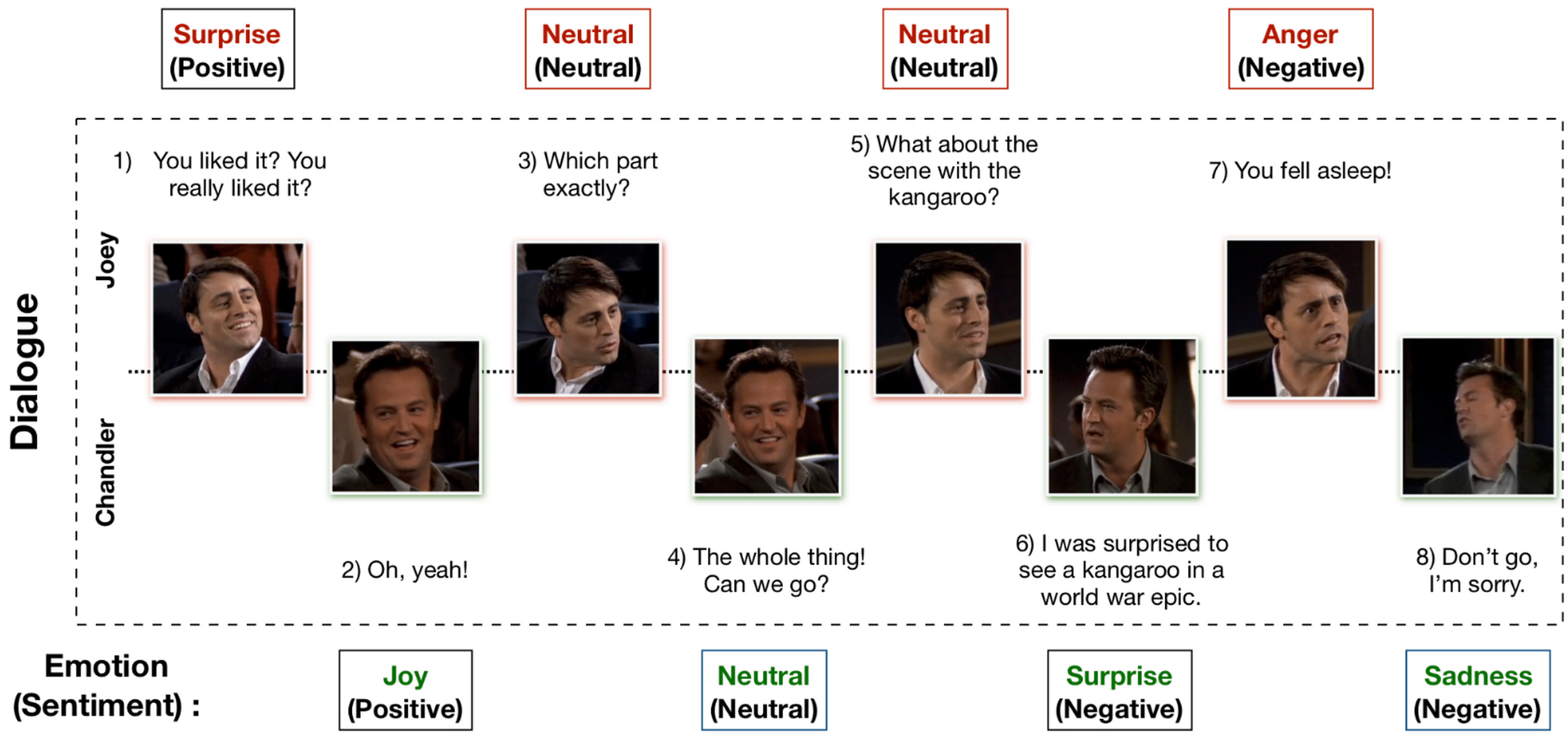


Similar bilinear models

Cross-Modality Relevance for Reasoning on Language and Vision, ACL 2020

Emotions are Often Context Dependent

NEW paper



"COSMIC: COmmonSense knowledge for eMotion Identification in Conversations", Findings of EMNLP 2020

Commonsense and Emotion Recognition

Proposed approach (COSMIC):

For each utterance, try to infer

speaker's intention
effect on the speaker/listener
reaction of the speaker/listener

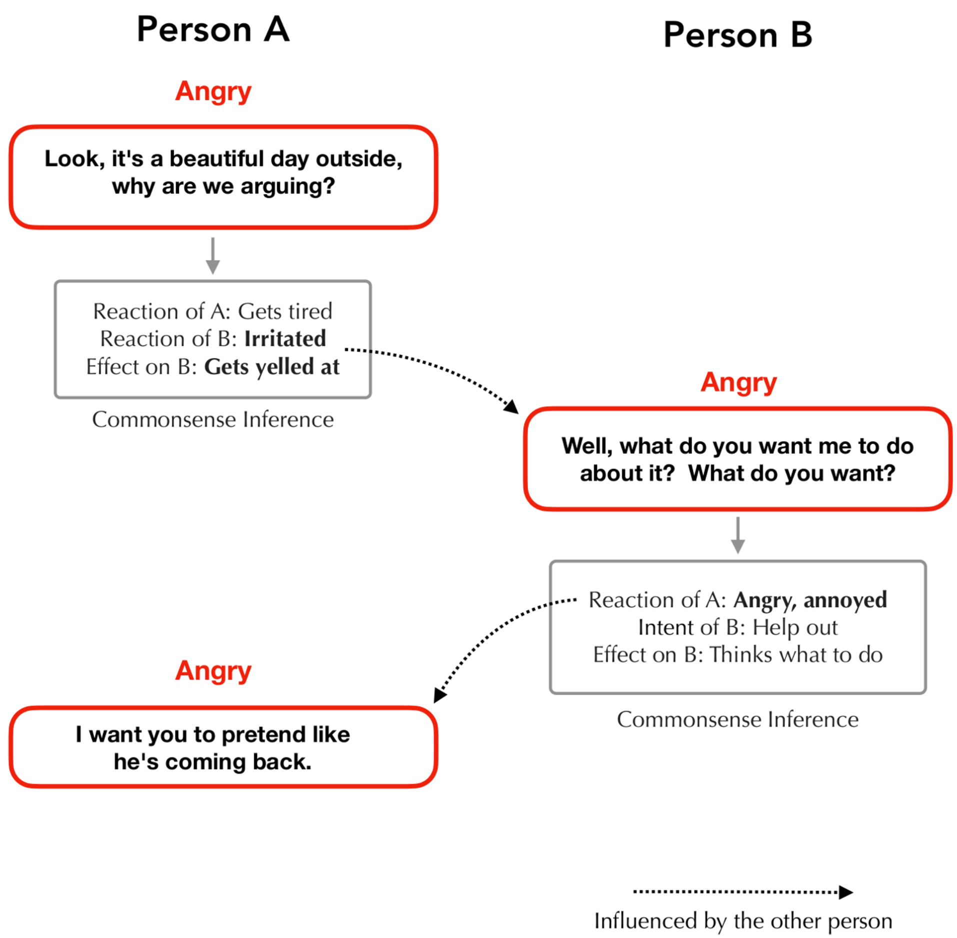
Example: "Person X gives Person Y a compliment"

→ Intend of X: "X wanted to be nice"

→ Reaction of Y: "Y will feel flattered"

"COSMIC: COmmonSense knowledge for eMotion Identification in Conversations", Findings of EMNLP 2020

Commonsense and emotion recognition



"COSMIC: COmmonSense knowledge for eMotion Identification in Conversations", Findings of EMNLP 2020